一种新的基于局部保持投影的高维数据聚类成员构造方法

周静波 殷 俊 金 忠
(南京理工大学计算机科学与技术学院 南京 210094)

摘 要 研究在高维数据中如何产生聚类成员,并提出一种新的构造聚类成员的方法。为解决高维数据的维度对构造成员带来的影响,新的构造方法在构造聚类成员之前利用局部保持投影先对高维数据进行维度约减,然后在约减后的子空间中用随机投影结合 K 均值方法构造聚类成员。最后讨论了局部保持投影子空间维度的选取。实验表明,新方法得到的结果要明显优于已有的主分量分析结合下采样方法和简单的随机投影方法。

关键词 聚类融合,维度约减,局部保持投影,随机投影
中图法分类号 TP391.4 **文献标识码** A

New Ensemble Constructor Based on Locality Preserving Projection for High Dimensional Clustering

ZHOU Jing-bo YIN Jun JIN Zhong
(School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract This paper studied how to construct cluster ensembles for high dimensional data and proposed a new ensemble constructor. To ameliorate the effect caused by high dimensionality, the proposed method used Locality Preserving Projections(LPP) to reduce the dimensionality before constructing ensembles. Then constructed ensembles based on random projection combined with K means in LPP subspace. Finally, we discussed how to choose the dimensionality of LPP subspace. The experiments show that ensembles generated by new algorithms perform better than those by Principal Component Analysis with subsampling(PCASS) and simple Random Projection(RP) that was proposed before.

Keywords Cluster ensembles, Dimension reduction, Locality preserving projections, Random projection

1 引言

在机器学习、模式识别和计算机视觉等领域中,数据聚类研究一直是一个非常活跃并且重要的课题[1],其目的是通过相似性度量,将数据集合划分为若干有意义的子集,从而发现隐藏的数据内部结构。由于很多聚类算法都利用两个数据点之间的距离,作为衡量数据点之间的相似性度量,因此,高维空间中聚类算法的性能直接受维度的影响。

高维数据对聚类算法提出了两大挑战[2]:一是高维数据分布大多比较稀疏,使得聚类算法检测数据内在结构变得十分困难;二是高维数据存在大量冗余特征,容易误导聚类算法的执行。为了克服由高维引起的问题,聚类算法常常与维度约减方法结合使用。

维度约减方法有很多,具有代表性的有主分量分析(PCA)[3]、局部保持投影(LPP)[4]、局部线性嵌入[5](LLE)等。这些方法都参照不同的准则,来求取高维数据集的低维表示。而根据这些准则得到的最优投影不一定能保持数据的内在结构。随机投影(RP)被证明在处理高维数据集时具有较大优势[2,13]。随机投影方法具有不稳定性[2],为得到较好的结果,一般多次采用随机投影得到多个数据的低维表示,然后对这些降维后的数据进行聚类,得到一系列聚类结果,这些聚类结果称为聚类成员。最后希望综合这些聚类成员,得到最终的结果能够体现原始数据的内在结构。聚类融合[6]算法能很好地解决这个问题。

在聚类融合过程中,只有聚类成员之间具有较大差异性时,将它们结合才能有效提高融合结果[7]。Fred 提出了 K 均值算法,利用不同的初始点产生不同聚类成员的方法[8]。Alexander Topchy 将数据随机投影到一维或多维坐标上,然后运用 K 均值聚类算法产生不同的聚类成员[9]。X. Z. Fern 采用随机投影方法结合 EM 算法得到不同聚类成员[2]。Roberto Avogadri 则先将数据嵌入到一个低维空间,使得低维嵌入保持原始数据集中数据点之间的距离特性,然后使用随机投影方法结合聚类算法来得到不同的聚类成员[11]。这些方法都能产生具有差异性的聚类成员,但是处理高维数据时,它们依旧受维度的影响。

本文主要讨论在高维条件下如何更加有效地产生聚类成员。提出了基于局部保持投影的构造聚类成员方法:利用局部保持投影方法得到高维数据的低维表示,然后采用随机投影方法得到不同的聚类成员。通过融合聚类成员后得到的性能比较,以及差异度-质量分析,都能说明这种方法的优越性。

2 相关工作

问题描述:给定含有 n 个数据点的数据集 $X=\{X_1, X_2,$

到稿日期:2010-10-29 返修日期:2010-12-30 本文受国家自然科学基金(60632050,60873151)资助。
周静波(1983-),男,博士生,主要研究领域为数据聚类、图像分割,E-mail:zhoujingbo2006@yahoo.com.cn;殷 俊(1984-),男,博士生,主要研究领域为模式识别、人脸识别、稀疏表示;金 忠(1961-),男,博士,教授,主要研究领域为图像分析、机器学习、计算机视觉、人脸识别。

…，$X_n\}$，采用某种方法得到 r 个聚类成员，表示为 $\Pi=\{\pi^1,\pi^2,\cdots,\pi^r\}$。其中 π^i 为第 i 次聚类得到的结果。π^i 将数据集 X 划分为互不相连的 K^i 个簇，即 $\pi^i=\{c_1^i,c_2^i,\cdots,c_{K^i}^i\}$，其中 $\bigcup_k c_k^i=X$。融合 π^i 的结果，希望能体现 X 的内在结构。

下面，先介绍两类已有的构造聚类成员的方法：主成分分析和随机二次抽样相结合的方法与随机投影方法。

2.1 主成分分析和随机二次抽样相结合方法

主成分分析(PCA)是一种应用很广泛的维度约减技术。给定一个数据集，PCA 选择几个方向将其投影到低维空间，使得到的低维表示数据方差最大。为了利用 PCA 获得不同的聚类成员，将它与随机二次抽样相结合。这种构造聚类成员的方法称为 PCASS[14]。

给定一个高维数据集 $X\in R^{n\times d}$，首先用 PCA 将数据集的维度由 d 降到 $l(l<<d)$ 维。然后对得到的新数据集采用无放回抽样获取新数据集的一个子集 $X'\in R^{m\times l}(m<n)$，对 X' 进行聚类，得到聚类结果。那些没有被抽样的数据点则根据 Euclidean 距离归入到离它最近的簇中。

PCASS 首先利用 PCA 降低数据集的维度，以此能够降低计算复杂度、节省运行时间，但这以损失大量信息为代价，这些信息包括大量噪声和有利于聚类的信息。PCASS 构造聚类成员时，有利于聚类的这部分信息一旦失去，是不可能再利用的。所以 PCASS 具有这样一个特点：在利用 PCA 降维时，假如有利于聚类的信息得以保留，PCASS 产生每一个聚类成员时都会利用到这部分信息，因此得到的聚类成员的质量较好，融合这些聚类成员可以得出令人满意的结果。相反，如果 PCA 降维时失去了这部分信息，产生的聚类成员质量很低，融合这些成员的最终结果不可能体现原始数据集的实际结构。

2.2 基于随机投影的方法

基于随机投影(Random Projection，RP)方法的基本思想来自于 JL 定理。RP 方法经常被用来进行高维数据的维数约简，同时也用来对文本、图像和音频数据降维、高维空间的最近邻搜索等。Fern 研究用 RP 约简数据进行聚类[2]。陈华辉等研究基于 RP 的并行数据流聚类[15]。

RP 首先需要产生一个随机投影矩阵 $R\in \mathrm{R}^{d\times l}(l<d)$，$R$ 由独立同分布的正态分布向量组成，并将其列归一化得到。给定高维数据集 $X\in \mathrm{R}^{n\times d}$，$X'=X\times R$，然后用标准的 K 均值算法对新数据集 X' 聚类，就可获得一个聚类成员。重复上述过程，由于每次产生的随机矩阵不同，因此得到的聚类成员也会不一样。用随机投影来产生聚类成员的方法称为基于 RP 的方法。

RP 不同于 PCASS。它不是依照某一准则来产生投影轴，而是随机产生投影轴。在构造聚类成员时，即使某一次数据集投影在产生的投影轴上损失掉了有利于聚类的这部分信息，也可以在下次投影时利用到这部分信息。换句话说，RP 得到的聚类成员之间的差异程度要比 PCASS 得到的聚类成员之间的差异程度大。所以，RP 一般要优于 PCASS。

RP 在处理高维数据集时也会出现如下的两个问题：

(1)当数据集维度很高而聚类成员个数有限时，RP 产生的投影轴保留有利于聚类的信息的概率很低；

(2)当聚类成员数较大时，RP 虽然在投影时有可能保留了有利于聚类的信息，但是利用到这部分信息的聚类成员所占比例很小，甚至可以忽略不计。通过融合这些聚类成员，得到的结果依然不能体现数据集的原始结构。

2.3 融合方法

在产生聚类成员之后，需要将聚类成员融合，得到最终结果。已有大量文献对融合这些聚类成员的方法进行了深刻研究[6,8-10]，不同的融合方法体现在不同设计的共识函数上。实验中采用 A. Strehl 提出的 CSPA 方法[6]来合并聚类成员。

CSPA 构造了一个这样的图，它映射了给定的数据集 X 中任意两个数据点之间的关系。给定一个聚类成员 $\Pi=\{\pi^1,\pi^2,\cdots,\pi^R\}$，CSPA 构造一个全连通图 $G=(V,W)$，其中

(1)V 是含有 n 个顶点的数据集，每个顶点代表数据集中的一个数据点。

(2)W 是一个相似矩阵

$$W(i,j)=\frac{1}{R}\sum_{r=1}^{R}I(g_r(X_i)=g_r(X_j)),I(g_r(X_i)=g_r(X_j))$$
$$=\begin{cases}1, & \text{if } g_r(X_i)=g_r(X_j)\\ 0, & \text{otherwise}\end{cases}$$

式中，$g_r(\cdot)$表示取一个数据点并返回 π^r 时它所属的类别。$W(i,j)$度量数据点 i 和数据点 j 在给定的聚类成员中被聚到同一个簇中的频率。

然后用基于图论的聚类算法 METIS 算法对 G 进行聚类，得到最终的聚类结果。

3 基于局部保持投影和随机投影结合的方法

M. Ester 提到利用局部邻域特征更有利于聚类[12]。为了克服 RP 存在的两个问题，我们提出了这样一种方法：首先应用维度约减方法对数据集降维，降维后的数据集应最大程度保留原始空间局部邻域信息，然后通过 RP 方法结合 K 均值算法来产生聚类成员。

3.1 局部保持投影(Locality Preserving Projection，LPP)

局部保持投影算法[4]是近年来提出的维度约减算法，最初目的是用于非线性流形的学习和分析。

LPP 算法是基于光谱理论的一种方法，它能够保持原始数据的局部结构信息。假设 $X=\{x_1,x_2,\cdots,x_n\}$是原始数据所构成的向量集，且 $x_i\in \mathrm{R}^d$，$1\leqslant i\leqslant n$。LPP 寻找一个转换矩阵 p 并把高维数据 X 投影到一个低维子空间 Y 上，且 $Y\subset \mathrm{R}^l$，$l<<d$，即 $Y=p^TX$。同时，Y 保持了原始数据 X 的局部结构。

设最优转换矩阵 P 由基向量集$\{p_1,p_2,\cdots,p_l\}$组成。这些向量可以通过解决以下的目标函数而得到：

$$p=\arg\min_{p}\frac{p^TXLX^Tp}{p^TXDX^Tp} \tag{1}$$

式中，$L=D-S$ 是拉普拉斯矩阵，$D_{ii}=\sum_j W_{ij}$，W 可以通过邻接图构造。如果 x_i 是 x_j 的 k 近邻或 x_j 是 x_i 的 k 近邻，那么 $W_{ij}=\exp(-\|x_i-x_j\|^2/\beta)$，否则 $W_{ij}=0$。

对角阵 D 中每个元素表示聚类数据的重要性，于是可以加入如下约束：

$$p^TXDX^Tp=1 \tag{2}$$

则式(1)可以转化为：

$$p=\arg\min_{p^TXDX^Tp=1}p^TXLX^Tp \tag{3}$$

根据拉格朗日定理，设辅助函数为 $F(\lambda,p)=p^TXLX^Tp-\lambda(p^TXDX^Tp-1)$，分别对 λ,p 求导：

$$\begin{cases}2XLX^Tp-2\lambda XDX^Tp=0\\ p^TXDX^Tp-1=0\end{cases} \tag{4}$$

于是，式(1)可以转化为求如下特征值问题：

$$XLX^{\mathrm{T}}p=\lambda XDX^{\mathrm{T}}p \tag{5}$$

那么基向量 $P=\{p_1,p_2,\cdots,p_l\}$ 是矩阵 $(XDX^{\mathrm{T}})^{-1}XLX^{\mathrm{T}}$ 的前 l 个最小特征值所对应的特征向量。因为矩阵 $(XDX^{\mathrm{T}})^{-1}XLX^{\mathrm{T}}$ 不是对称的，所以 $\{p_1,p_2,\cdots,p_l\}$ 不一定是正交的，这使得重构数据很困难。为解决这个问题，在实验中利用 Cai Deng 提出的方法[4]对 $P=\{p_1,p_2,\cdots,p_l\}$ 进行正交化，得到一组正交的基向量：

$$W_{\mathrm{LPP}}=\{y_1,y_2,\cdots,y_l\} \tag{6}$$

LPP 能够较好地保留原始数据的局部结构，对于类间距离较远的数据能找到较好的投影方向。特别地，LPP 对于类内有多个聚类的问题也能得到较好结果。然而，选择投影向量个数的多少，也即新空间的维度对 LPP 的性能影响较大。在 4.5 节，我们专门讨论了 LPP 维度 d 的选择方法。

3.2 结合 LPP 与 RP 方法产生聚类成员算法(LPPRP)

利用 LPP 保持局部邻域的特性，我们希望得到的低维数据能够最大程度保持原始数据的结构信息。为了得到差异性较大的聚类成员，算法将 LPP 与 RP 相结合。具体算法过程如表 1 所列。

表 1 结合 LPP 与 RP 方法产生聚类成员算法(LPPRP)

算法：LPPRP
输入：样本数据 X，投影向量数 l，聚类成员数 r，聚类数 k
输出：r 个聚类成员
Step 1 根据式(6)得到数据的低维投影 $X'=X\times W_{\mathrm{LPP}}$
Step 2 重复下述步骤 r 次
(a)根据 1.2 节计算随机投影矩阵 R，$X''=X'\times R$，进一步降低维度
(b)使用 K 均值算法将数据 X″聚成 k 类
Step 3 返回 r 个聚类成员 $\Pi=\{\pi^1,\pi^2,\cdots,\pi^r\}$

通过新算法得到 r 个聚类成员之后，利用 CSPA 这些成员融合成一个最终的结果。可以看到，LPPRP 在 RP 之前增加了一个 LPP 降维的过程。在 LPP 降维后的子空间中通过 RP 产生聚类成员，增加了原始空间中有利于聚类的信息利用率。

(1)在产生单个聚类成员时，利用到这部分信息的几率增大；

(2)在产生大量聚类成员时，利用到这部分信息的聚类成员所占比例增大。因此，处理高维数据集时，LPPRP 的性能较 RP 优越。

4 实验与分析

4.1 数据集描述

在实验中使用了 4 组数据，表 2 简单描述了这些数据集。其中 ISOLET6 数据集是 UCI 语音字符识别数据集的子集，选择了 A,D,F,H,K,Y 字符数据。MFEAT 数据集来自 UCI 多维特征数据集，原始数据集有很多种不同的表示方法，这里使用傅里叶系数表示的数据集。表 3 描述了这些数据集的特性并列出了每个数据集实验过程中选择的参数。

表 2 数据集描述

数据集名称	描述	数据源
CHART	综合生成的控制图时间序列	UCI KDD 档案
ISOLET6	语音字符数据集(6 个字符)	UCI ML 档案
MFEAT	手写体数字傅里叶系数表示	UCI ML 档案
ORL face	人脸数据库	AT&T

表 3 数据集概要

	CHART	ISOLET6	MFEAT	ORL face
样本数	600	1800	2000	400
类别数	6	6	10	40
初始维数	60	617	76	4096
d1	30	60	30	60
d2	10	30	10	30
k	10	10	20	50

注：d1 表示经过 LPP 约减后的空间维数，d2 表示 PCA，RP 约减后的空间维数，k 是运行 K 均值时选择的聚类数。

4.2 评价准则

聚类算法的性能评价一直是一个具有挑战性的问题。本文使用规范化互信息(NMI)[6]作为聚类的评价方法。如果 C 是样本聚类以后的类标号，Y 是样本实际的类标号，则规范化的互信息表示为：

$$\mathrm{NMI}(C,Y)=\frac{I(C;Y)}{(H(C)+H(Y))/2} \tag{7}$$

式中，$I(C;Y)=H(Y)-H(Y|C)$ 是 C 和 Y 之间的互信息，$H(Y)$ 是 Y 的香农熵，$H(Y|C)$ 是给定 C 的条件下，Y 的条件熵。NMI 值范围在 0 和 1 之间，NMI 值越大，聚类的性能就越好。

4.3 实验分析

图 1 是在上述 4 个数据集上分别使用 PCASS，RP，LPPRP 方法获取聚类成员后合并得到的一个聚类解与数据原有标签的 NMI 值。为合理评价各种方法，将每一种构造聚类成员方法在每一个数据集上运行 10 次，取它们的平均值作为最后的结果。图 1 中的 x 轴表示聚类成员的个数，其值选择为 1，10，20，…，100；y 轴表示对应聚类成员数的 NMI 值。在实验中，PCASS 随机下采样率为 70%，LPP 选择 8 个近邻点构造邻域图，其他参数参照 4.1 节表 3 设置。

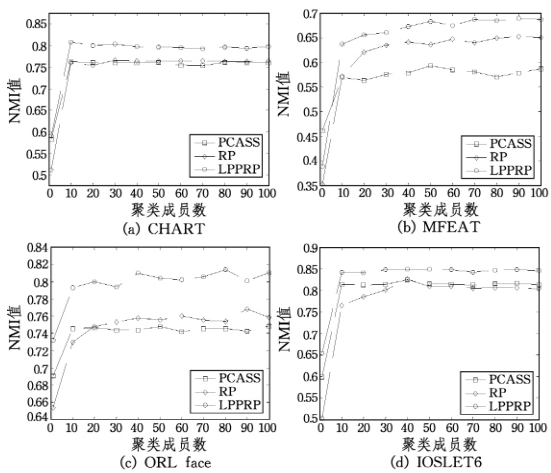

图 1 3 类算法在不同数据集上的性能

从结果中可以看到，LPPRP 要优于 RP 和 PCASS。当样本原始维数较小时，新方法提高的效果相对不是很明显。当原始样本维数较大时，LPPRP 得到的效果要明显好于 RP 和 PCASS。而 RP 整体优于 PCASS 结果，如图 1(MFE-AT)，(ORL face)所示。如果 PCA 降维保留了有利于聚类的信息，则 PCASS 比 RP 好，如图 1(ISOLET6)所示。

利用 LPP 对高维数据集进行降维，得到高维数据的低维表示。一般来说，这个低维表示保持了数据分布的邻域信息。所以 LPP 降维过程中更容易保留原始空间中有利于聚类的信息。在低维条件下，使用 RP 可以提高这部分信息的利用率。因此，LPPRP 构造的聚类成员的质量要优于 RP 和

PCASS，这跟实验结果是一致的。

为了更好地评价 PCASS，RP，LPPRP 这 3 种方法的优劣，下面给出它们的差异度-质量分析[2]。

4.4 差异度-质量分析(diversity-quality，d-q)

给定一组聚类成员，计算所有聚类成员两两之间的 NMI 值。每个 NMI 值都衡量了两个聚类成员之间的差异度。质量度量则是计算每个聚类成员与数据集实际标签之间的 NMI 值。注意到，当两个聚类成员之间的 NMI 值为 0 时，它们之间的差异度最大；相反地，聚类成员与数据集原有标签的 NMI 值越大，这组聚类成员的质量越好。所以，一组好的聚类成员须同时满足较大的差异度和较好的聚类质量。

下面比较 RP，PCASS，LPPRP 3 种方法所产生的聚类成员差异度和质量。对每一种方法，取聚类成员个数为 50 个，其他参数设置参照表 3。根据计算得到表 2 中 4 个数据集的差异度-质量分析，如图 2 所示。图中点由两个值(d,q)表示。d 表示聚类质量，通过计算聚类成员与数据集实际标签的 NMI 值得到；q 表示聚类成员之间的差异度，通过计算一个聚类成员与其他聚类成员的 NMI 值的平均值得到。聚类成员要同时满足较大的差异性和较好的质量，它的 $d-q$ 值应该落在右下方的区域，或者靠近这个区域。

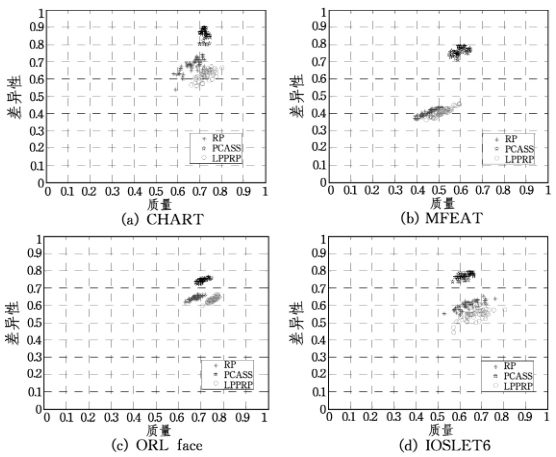

图 2 3 种算法的差异度-质量分析

从图 2 可以看出，要同时满足最大的差异度和较发的质量是不可能的。因为聚类成员只有 50 个，要保持聚类成员的最大差异度，则聚类成员两两之间的 NMI 值都会比较小。当成员中某一个与数据集原始标签之间的 NMI 值较大时，其他成员与数据集原始标签之间的 NMI 都会比较小，所以计算得到聚类成员的质量也会变小。下面分别分析 PCASS，RP，LPPRP 方法的差异度和质量度量。PCASS 得到的聚类成员差异度是最小的，说明 PCASS 得到的聚类成员具有很高的相似性。融合这些具有高相似性的聚类成员后的性能并没有提高。RP 方法产生的聚类成员相对于 PCASS 方法差异度大，质量小。从图 1 中可以看到，融合后前者的性能却比后者好。对比 RP 和 LPPRP 方法发现，RP 方法得到的聚类成员的差异度比 LPPRP 的要小，而成员质量却比 LPPRP 的差。所以我们认为 RP 方法在处理高维数据集时不如 LPPRP 好。

4.5 参数讨论

利用 LPP 对数据集降维时，需要确定低维表示的维度 d。为了说明问题，选取 d 的范围为$[40,200]$，其他参数设置参照表 3，在数据集 ISOLET6 上运行多次后得到的结果如图 3 所示。LPPRP 选取 $d=40$，$d=80$，$d=160$，$d=200$ 时算法的性能都不如 $d=120$ 好。可以得到这样一个结论：当维度 d 取值较小时，容易损失掉有利于聚类的信息；而 d 取值较大时，虽然可能保留更多的信息，但是利用到有利于聚类的那部分信息的概率也随之下降。

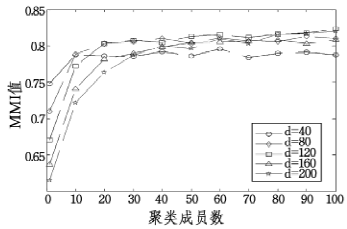

图 3 LPPRP 在 ISOLET6 数据集上选择不同 d 得到的算法性能

另外，当聚类成员数为 1 时，可以看到当 $d=200$ 时算法的性能最差，当 $d=40$ 时性能最好。因此，当 d 取值较小时，局部保持投影后的信息被压缩在一个很小的空间，利用 RP 产生聚类成员时应用到这部分信息的概率很大。所以，当产生较少聚类成员时，d 应当取较小值。

结束语 本文研究了在高维数据集中如何利用维度约减方法降低数据集维度，并产生有效的聚类成员的方法。同时提出了通过 LPP 维度约减方法对高维数据降维，并在降维过程中保持局部邻域的信息；其次利用随机投影法来产生聚类成员；最后，将得到的聚类成员合并，并产生一个最终解。实验结果表明，新方法产生的聚类成员要比 RP 和 PCASS 方法好。

去除数据空间中的冗余特征，使保留的信息能够揭示高维数据的结构，一直是高维数据聚类的难点。在高维数据聚类时，如何选择维度约减方法，使得高维数据集中有利于聚类的信息能够尽可能保留下来，并且在聚类时最大程度利用到这部分信息，将是我们下一阶段工作的方向。

参 考 文 献

[1] Jain A K. Data Clustering：50 Years Beyond K-Means[J]. Pattern Recognition Letters，2010，31(8)：651-666

[2] Fern X Z，Brodley C E. Random Projection For High Dimensional Data Clustering：A Cluster Ensemble Approach[C]// Proceedings of the 20th International Conference on Machine Learning. Washington DC，2003：186-193

[3] Turk M，Pentland A P. Face Recognition Using Eigenfaces[C]// IEEE Conference on Computer Vision and Pattern Recognition. Maui Marriott，Hawaii，1991：586-591

[4] Deng Cai，et al. Orthogonal Laplacianfaces for Face Recognition [J]. IEEE Transactions on Image Processing，2006，15(11)：3608-3614

[5] Roweis S T，Saul L K. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science，2000，290(5500)：2323-2326

[6] Strehl A，Ghost J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions[J]. Journal of Machine Learning Research，2002，3：583-617

[7] 罗会兰，孔繁盛，李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报，2007，30(8)：1315-1325

[8] Fred A L，Jain A K. Combining Multiple Clusterings Using Evidence Accumulation[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence，2005：835-850

[9] Topchy A，Jain A K. Clustering Ensembles：Models of Consensus and Weak Partitions[J]. IEEE Transaction on Pattern Anal-

ysis and Machine Intelligence,2005,27(6):1866-1881

[10] Ayad H G,Kamel M S. Cumulative Voting Consensus Method for Partitions with A Variable Number of Clusters[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2008,30(1):160-173

[11] Avogadri R, Valentini G. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis[J]. Artificial Intelligence in Medicine,2009,45(2/3):173-183

[12] Ester M,Kriegel H-P,et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// The ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Portland,Oregon,1996:226-231

[13] Azimi J,Fern X. Adaptive Cluster Ensemble Selection[C]// Proceedings of International Joint Conference on Artificial Intelligence(IJCAI). California,2009:992-997

[14] Fern X Z,Brodley C E. Cluster Ensembles for High Dimensional Clustering:An Empirical Study[EB/OL]. http://Web. engr. oregonstate. edu/~xfern/clustensem. pdf,2010-02-20

[15] 陈华辉,施伯乐. 基于随机投影的并行数据流聚类方法[J]. 模式识别与人工智能,2009,22:113-122

(上接第 154 页)

全局文档空间中通过最小化损失函数来实现用本体对文档的语义标注。基于这个模型,本文实现了一个针对非结构化文档的语义标注工具。通过这个工具所做的实验表明,该方法能有效地对互联网中大量以网页等形式存在的、质量良莠不齐的多种类文档知识资源进行有效的自动语义标注。这不仅对于质量较好的文档标注结果好,对于质量差的文档资源也能取得让人接受的结果,即该方法对文档质量的抗干扰能力较强。另一方面,本体质量的优劣同样对标注结果有重要影响。实验也同样表明,该方法受本体质量变化的影响相对较小。

通过本文的工作我们认识到,网络中文本资源质量差异很大,其中质量不好的文本资源占相当的比重。因而,我们认为,如果不能对这些质量相对较差的资源进行有效的标注,会造成网络资源丢失或浪费。另一方面,文档所描述的知识关系往往稀疏地分布在文档的不同段落中,且段落与段落之间本身有一定的关系。段落之间关系密切,则其描述的主题相同或相近的概率较大。本文的方法对知识分布稀疏的段落处理能力较好,但对关系密切段落之间还没有相应的对策,这将是我们下一步需要研究的工作。

参 考 文 献

[1] Uschold M. Converting an Informal Ontology into Ontolingua [C]// Proceedings of the Workshop on Ontological Engineering held in conjunction with ECAI 96. Budapest,March 1996

[2] Dill S,Eiron N,Gibson D,et al. A Case for Automated Large Scale Semantic Annotation[J]. Journal of Web Sematics,2003 (1):115-132

[3] Handschuh S, Staab S, Ciravegna F. S-cream Semi-automatic Creation of Meta-data[C]// Gómez-Pérez A, Benjamins V R, eds. 13th Intl. Conf. on Knowledge Engineering and Knowledge Management Ontologies and the Semantic Web. LNCS, Vol. 2473. Berlin Heidelberg New York:Springer Verlag,2002:358-372

[4] Kiryakov A,Popov B,Terziev I,et al. Semantic Annotation,Indexing,and Retrieval[J]. Journal of Web Sematics 2,2004(1): 47-49

[5] Popov B, Kiryakov A, Ognyanoff D, et al. KIM: A Semantic Platform for Information Extaction and Retrieval[J]. Journal of Natural Language Engineering, Cambridge University Press, 2004(3/4):375-392

[6] Ciravegna F,Wilks Y. Designing adaptive informationextraction for the Semantic Web in amilcare, HandschuhS[C]// Staab S, ed. Annotation for the Semantic Web,Frontiers in Artificial Intelligence and Applications. Amsterdam:IOS Press,2003:112-127

[7] Handsehuh S,Staab S,Maedche A. CREAM:Creating relational metadata with a component-based, ontology-driven annotation framework[C]// Proc. of the 1st Int'l Conf on Knowledge Capture. New York:ACM,2001:76-83

[8] Gregory G. Use of syntactic context to produce term association lists for text retrieval[C]// Belkin N,Ingwersen P,Pejtersen A M,eds. Proc. of the 15th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Copenhagen: ACM Press,1992:89-97

[9] Chang Y,Ounis I,Kim M. Query reformulation using automatically generated query concepts from a document space[J]. Information Processing and Management,2006,42:453-468

[10] Yuan L,Li Z H,Chen S L. Ontology-based annotation for deep Web data[J]. Journal of Software,2008,19(2):237-245

[11] Ma An-xiang,Zhang Bin,Gao Ke-ning,et al. Deep Web Data Extraction Based on Result Pattern[J]. Journal of Computer Research and Development,2009 46(2):280-288

[12] Gardent C,Parmentier Y. SemTAG:a platform for specifying tree adjoining grammars and performing TAG-based semantic construction[C]// Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. 2007:13-16

[13] Tenier S,Toussaint Y,Napoli A,et al. Instantiation of relations for semantic annotation[C]// Proc. of Web Intelligence 2006. Los Alam itos:IEEE Computer Society. 2006:463-472

[14] Xu J X,Croft W B. Improving the effectiveness of information retrieval with local context analysis[J]. ACM Trans. on Information Systems,2000,18(1):79-112

[15] Zhang M,Song R H,Ma S P. Document Refinternet based on semantic query expansion[J]. Chinese Journal of Computers, 2004,27(10):1395-1401

[16] Jang M G,Myaeng S H,Park S Y. Using mutual information to resolve query translation ambiguities and query term weighting [C]// Dale R,Chuch K. eds. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. College Park: Association for Computational Linguistics,1999:223-229

[17] Gao J F,Zhou M,Nie J Y,et al. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations[C]// Järvelin K,Chairs P,Baeza-Yates R,et al. eds. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere: ACM Press,2002:183-190